A digitisation landscape: the IMPACT Centre of Competence dataset

Gustavo Candela, Isabel Martínez, Katrien Depuydt, Tomasz Parkola, Sally Chambers & Błażej Betański

(University of Alicante, University of Alicante, Instituut voor de Nederlandse Taal, Poznań Supercomputing and Networking Center, Royal Library of Belgium & Poznań Supercomputing and Networking Center)

Keywords: Text recognition, Optical character recognition, Digitisation, Collections as data, Reuse

GLAM and library initiatives have provided digital collections for decades. Recent initiatives such as Collection as data (Padilla, 2019) and GLAM Labs (Mahey, 2019) promote the reuse of the content using computational methods. In this sense, community-led approaches have encouraged institutions to adopt new techniques in terms of digitisation. For example, the IMPACT Centre of Competence[1] has promoted the digitisation of content in cultural heritage organisations as well as to further advance the state-of-the-art in the field of document imaging, language technology, quality assessment and the processing of historical text.

One of the services the centre offers is access to the ground truth dataset developed in the IMPACT project. The IMPACT dataset contains more than half a million representative text-based images compiled by a number of major European libraries. Covering texts from as early as 1500, and containing material from newspapers, books, pamphlets and typewritten notes, the dataset is an invaluable resource for future research into imaging technology, OCR and language enrichment. A part of this dataset (ca. 50,000 items) has ground truth data associated with it; a transcription of the image content including layout information. The ground truth provided is stored and exchanged in xml in the Page Analysis and Ground-truth Elements (PAGE) format. In this context, a new web platform has been released based on the DInGO/dLibra software tool developed and hosted by the Poznan Supercomputing and Networking Center (PSNC). Instead of providing individual downloadable files, and according to best practices (Candela, 2023), the tool was developed through an iterative process and enables browsing, searching and downloading of the main documents of the dataset.

IMPACT seeks to provide innovative examples of use of the dataset, to encourage further engagement with the dataset. Some examples include: i) the use of the ground truth for OCR quality evaluation; ii) to provide reproducible and open code based on a collection of Jupyter Notebooks to show users how to access and reuse the content; and iii) to create/adjust language models based on the text provided by the dataset.

In order to engage with the digitisation community and GLAM institutions, the new tool enables browsing the content of the IMPACT dataset. We will expand the repository by also inviting new institutions and researchers to make their ground truth data and images available. Future work includes the addition of the platform to DH platforms such as the Social Sciences and Humanities Open Marketplace.

[1] https://www.digitisation.eu/, IMPACT is a not for profit organisation with the mission to make the digitisation of text “better, faster, cheaper” and to further advance the state-of-the-art in the field of document imaging, language technology and the processing of historical text.

References

Padilla, T., et al. (2019). Final Report --- Always Already Computational: Collections as Data (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3152935

Mahey, M., et al. (2019) Open a GLAM Lab. Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23-27 September 2019. https://glamlabs.io/books/open-a-glam-lab/

Candela, G., et al. (2023), "A checklist to publish collections as data in GLAM institutions", Global Knowledge, Memory and Communication, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/GKMC-06-2023-0195